JMLR: Workshop and Conference Proceedings 1:??–??, 2014 ECML/PKDD Workshop on Statistically Sound Data Mining

# Look before you leap: Some insights into learner evaluation with cross-validation

**Gitte Vanwinckelen** GITTE.VANWINCKELEN@CS.KULEUVEN.BE
**Hendrik Blockeel** HENDRIK.BLOCKEEL@CS.KULEUVEN.BE
*KU Leuven, Department of Computer Science*
*Celestijnenlaan 200A*
*3001 Heverlee, Belgium*

**Editor:** Editor's name

## Abstract

Machine learning is largely an experimental science, of which the evaluation of predictive models is an important aspect. These days, cross-validation is the most widely used method for this task. There are, however, a number of important points that should be taken into account when using this methodology. First, one should clearly state what they are trying to estimate. Namely, a distinction should be made between the evaluation of a *model* learned on a single dataset, and that of a *learner* trained on a random sample from a given data population. Each of these two questions requires a different statistical approach and should not be confused with each other. While this has been noted before, the literature on this topic is generally not very accessible. This paper tries to give an understandable overview of the statistical aspects of these two evaluation tasks. We also pose that because of the often limited availability of data, and the difficulty of selecting an appropriate statistical test, it is in some cases perhaps better to abstain from statistical testing, and instead focus on an interpretation of the immediate results.

**Keywords:** cross-validation, classification, learner evaluation

## 1. Introduction

Most machine learning tasks can be addressed using multiple alternative learning methods. Empirical performance evaluation plays an important role here. The behavior of all these methods is not always theoretically well-understood, and if it is, it is important to stress the real world implications. Therefore, almost all papers contain some form of performance evaluation, usually estimating the quality of models resulting from the machine learning effort (for instance, predictive performance), the computational effort required to obtain these models, or other performance criteria.[1]

For predictive models, a major criterion is usually the accuracy of the predictions, or more generally, the expected "loss", using a loss function that compares the predicted values with the correct ones. Much research in machine learning focuses on developing better learning methods, that is, methods that are more likely to return models with a lower expected loss.

1. In line with most machine learning literature, and somewhat at variance with the statistical literature, the term "model" here refers to the result of the learning effort (e.g., a specific decision tree), not to the type of model considered (e.g., "decision trees").

This goal statement is still somewhat vague, and can be interpreted in multiple ways. From the no-free-lunch theorems (**?**), we know that, averaged over all possible learning tasks, all learners perform equally well, so the goal only makes sense when the set of tasks is restricted to, for instance, a specific application domain. A specific learning task can be formalized using a single population. The task is then to learn a model for this population from a dataset sampled at random from it. The following two different versions of this task can then be distinguished.

1. Given a dataset $D$ from population $P$, and a set of learners, which learner learns from $D$ the most accurate model on $P$?

2. Given a population $P$, and a set of learners, which learner is expected to yield the most accurate model on $P$, when given a random sample of a particular size from $P$?

The first question is relevant for researchers who evaluate the learning algorithms using the same dataset that an end user will use to build a model. The second question is relevant when the end user's dataset is not available to the researcher.

Authors rarely clarify which of these two questions they try to answer when evaluating learning methods. This is often clear from the context. For instance, when testing a learning method on UCI datasets (**?**), one is clearly not interested in the models learned from these datasets, but in the behavior of the learner on other, similar learning problems, where "similar" is to be interpreted here as "learning from a dataset of similar size, sampled from a population with a distribution similar to that of the UCI dataset's population". [2] On the other hand, when learning predictive models from a given protein-protein interaction network, one may well be interested in the predictive quality of these specific models.

Not making the question explicit carries a risk. Both questions require a different approach and different statistical tests, and leaving the question implicit may obfuscate the fact that the wrong statistical methods are used.

In the statistical literature, the two questions are clearly distinguished, and studied separately. However, this literature is not always very accessible to the machine learning audience; relevant information is spread over many different articles that are often quite technical.

The goal of this article is to increase awareness in the machine learning community about the difference between the above two questions, to summarize the existing knowledge about this difference in an accessible manner, and to provide guidance on empirical evaluation to machine learning researchers.

The remainder of this work is organized as follows. We first define the general task of evaluating a predictive model (Section **??**). Next, we define how to measure the performance with the error as loss function, differentiating between model and learner evaluation (Section **??**). We then introduce cross-validation, which is typically used to get an estimate of the real error of a model (Section **??**). We define the cross-validation estimator as a stochastic function of the sample on which it is computed, and of how this sample

2. The qualification "of similar size" for the dataset is needed because the quality of a learned model depends on the size of the dataset from which it was learned (see, e.g., (**?**)), and the qualification of the distribution is needed because it is well-known that no learner can be optimal for all population distributions (**?**)

is partitioned. The expected difference between the cross-validation estimate and the true error is then quantified by its mean squared error (Section **??**). An accurate estimate of the mean squared error informs us how confident we can be about the conclusions from our learner evaluation. We therefore discuss the pitfalls of estimating this quantity (Section **??**). Finally, we supplement our theoretical discussion with two experiments. The first experiment demonstrates some aspects of evaluating a model or a learner with repeated cross-validation. The second experiment investigates the uncertainty about selecting the winning model when comparing two models with cross-validation (Section **??**).

## 2. Preliminaries

### 2.1. Learning task

We focus on the setting of learning predictive functions from examples of input-output pairs. In the following, $2^S$ denotes the power set of $S$ and $\mathcal{Y}^{\mathcal{X}}$ denotes the set of all functions from $\mathcal{X}$ to $\mathcal{Y}$. We formalize learning tasks as follows.

**Definition 1 (predictive learning)** *A predictive learning task is a tuple $(\mathcal{X}, \mathcal{Y}, p, T, C)$, where $\mathcal{X}$ is called the input space, $\mathcal{Y}$ is called the output space, $p$ is a probability distribution over $\mathcal{X} \times \mathcal{Y}$, $T \subseteq \mathcal{X} \times \mathcal{Y}$ is called the training set, and $C : \mathcal{Y}^{\mathcal{X}} \times \mathcal{P} \to \mathbb{R}$ (with $\mathcal{P}$ the set of all distributions over $\mathcal{X} \times \mathcal{Y}$) is some criterion to be optimized.*

The probability distribution $p$ is called the *population distribution*, or simply population. Without loss of generality, we assume from here on that $C$ is to be *minimized*.

**Definition 2 (learner)** *A learner $L$ is a function with signature $L : 2^{\mathcal{X} \times \mathcal{Y}} \to \mathcal{Y}^{\mathcal{X}}$.*

**Definition 3 (performance)** *Given a learning task $(\mathcal{X}, \mathcal{Y}, p, T, C)$, a learner $L_1$ has better performance than a learner $L_2$ if $C(L_1(T), p) < C(L_2(T), p)$.*

Note that, as the goal of predictive learning is to find a model that can make predictions for instances we have not seen before, the quality criterion for the resulting model is based on the population $p$, not on the training set $T$. Differently from $T$, however, $p$ is not known to the learner. It is often also not known to the researcher evaluating the method.

In the following, we assume that the output space $\mathcal{Y}$ is one-dimensional. If $\mathcal{Y}$ is a set of nominal values, the learning task is called classification; if $\mathcal{Y}$ is numerical, the task is called regression.

### 2.2. Error measures

Much of the relevant literature on the estimation of learning performance focuses on regression and classification tasks, and uses error as a performance measure. A few examples are: **?????**. We here focus on classification. We start with repeating some basic definitions used in that context.

In the following, $\mathbf{Pr}_{x \sim p}[C]$ denotes the probability of a boolean function $C$ of $x$ evaluating to true, and $\mathbf{E}_{x \sim p}[f(x)]$ denotes the expected value of $f(x)$, when $x$ is drawn according to $p$. For a set $T$, we use the notation $T \sim p$ to denote that all elements of $T$ are drawn independently according to $p$.

The concept of "error" can be defined on two levels: that of learners, and that of learned models (classifiers). The error of a classifier is defined as follows.

**Definition 4 (error)** *The error of a classifier $m$ is the probability of making an incorrect prediction for an instance drawn randomly from the population. That is,*

$$\varepsilon(m) = \mathbf{Pr}_{(\mathbf{x},y)\sim p}[m(\mathbf{x}) \neq y] \tag{1}$$

For learners, two types of error are typically distinguished: The conditional and the unconditional error (**?**, Chapter 7).

**Definition 5 (conditional error)** *The conditional error of a learner $L$ for a dataset $T$, denoted as $\varepsilon_c(L,T)$, is the error of the model that it learns from $T$.*

$$\varepsilon_c(L,T) = \varepsilon(L(T))., \text{ with } m_T = L(T). \tag{2}$$

**Definition 6 (unconditional error)** *The unconditional error of a learner $L$ at size $N$, denoted $\varepsilon_u(L,N)$, is the expected error of the model learned by $L$ from a random dataset of size $N$. It is the mean of the conditional error $\varepsilon_c(L,T)$ taken over all datasets of size $N$ that can be sampled from population $p$.*

$$\varepsilon_u(L,N) = \mathbf{E}_{\{T\sim p:|T|=N\}}[\varepsilon_c(L,T)]. \tag{3}$$

These two different types of error are clearly related to the two different questions mentioned in the introduction. The conditional error of a learner is relevant if the dataset $T$ used for the estimation is identical to the one that will be used by other researchers when learning predictive models for the population. The unconditional error is relevant if the dataset $T$ used for the estimation is representative for, but not identical to, the datasets that other researchers will use. It estimates the expected performance of the learner on similar datasets (that is: datasets of the same size sampled from the same distribution), rather than its performance on the given dataset.

In the remainder of this text, we focus on error as the criterion to be optimized, but it is clear that for any loss function, a distinction can be made between the conditional and unconditional version of that loss.

### 2.3. Cross-validation error estimator

As the population $p$ is usually unknown, the true error (conditional or unconditional) cannot be computed but must be estimated using the training set $T$. Many different estimation methods have been proposed, but by far the most popular estimators are based on cross-validation. It relies on the notion of empirical error:

**Definition 7 (Empirical error)** *The empirical error of a model $m$ on a set of instances $S$, denoted $e(m,S)$, is*

$$e(m,S) = \frac{|\{(\mathbf{x},y) \in S | m(\mathbf{x}) \neq y\}|}{|\{(\mathbf{x},y) \in S\}|}.$$

In *k-fold cross-validation*, a dataset $T$ is randomly divided into $k$ equally sized (up to one instance) non-overlapping subsets $T_i$, called folds. For each fold $T_i$, a training set $Tr_i$ is defined as $T \setminus T_i$, a model $m_i$ is learned from $Tr_i$, and $m_i$'s error is estimated on $T_i$. The mean of all these error estimates is returned as the final estimate.

**Definition 8** *The k-fold cross-validation estimator, denoted $CV_k(L,T)$, consists of partitioning $T$ in $k$ subsets $T_1, T_2, \ldots, T_k$ such that $|T_i| - |T_j| \leq 1 \; \forall \, i, j$, and computing*

$$CV_k(L,T) = \frac{1}{k} \sum_{i=1}^{k} e(L(T \setminus T_i), T_i)$$

If the number of folds $k$ equals the number of instances $|T|$ in the dataset, the resampling estimator is called *leave-one-out cross-validation*. This special case is usually studied separately.

**Definition 9** *The leave-one-out cross-validation estimator, denoted $CV_{|T|}(L,T)$, is*

$$CV_{|T|}(L,T) = \frac{1}{|T|} \sum_{i=1}^{|T|} e(L(T \setminus \{t_i\}), \{t_i\})$$

*with $T = \{t_1, t_2, \ldots, t_{|T|}\}$.*

Contrary to $CV_k$, which is a stochastic function, $CV_{|T|}$ is deterministic, as there is only one way to partition a set into singleton subsets.

*Repeated k-fold cross-validation* computes the mean of $n$ different $k$-fold cross-validations on the same dataset, each time using a different random partitioning.

**Definition 10** *The n-times repeated k-fold cross-validation estimator is*

$$RCV_{n,k}(L,T) = \frac{1}{n} \sum_{1}^{n} CV_k(L,T).$$

In practice, a stratified version of these estimators is often used. In stratified cross-validation, the random folds are chosen such that the class distribution in each fold is maximally similar to the class distribution in the whole set. Note that stratification is not possible in the case of leave-one-out cross-validation.

**Definition 11** *The stratified k-fold cross-validation estimator, denoted $SCV_k(L,T)$, consists of partitioning $T$ in $k$ equal-sized subsets $T_1, T_2, \ldots, T_k$ with class distributions equal to that of $T$, and computing*

$$SCV_k(L,T) = \frac{1}{k} \sum_{i=1}^{k} e(L(T \setminus T_i), T_i)$$

**Definition 12** *The n-times repeated stratified k-fold cross-validation estimator is*

$$RSCV_{n,k}(L,T) = \frac{1}{n} \sum_{1}^{n} SCV_k(L,T).$$

## 3. Estimator quality

Having introduced two population parameters, the conditional and the unconditional error, and the cross-validation estimator, we now investigate the quality of this estimator for either parameter.

When estimating a numerical population parameter $\varepsilon$ using an estimator $\hat{\varepsilon}$, the estimator's quality is typically expressed using its mean squared error, which can be decomposed in two components: bias and variance.

$$MSE(\hat{\varepsilon}, \varepsilon) = \mathbf{E}[(\hat{\varepsilon} - \varepsilon)^2] = \mathrm{Var}(\hat{\varepsilon}) + B^2(\hat{\varepsilon}, \varepsilon)$$

with

$$\mathrm{Var}(\hat{\varepsilon}) = \mathbf{E}[(\hat{\varepsilon} - \mathbf{E}[\hat{\varepsilon}])^2].$$

and

$$B(\hat{\varepsilon}, \varepsilon) = \mathbf{E}[\hat{\varepsilon}] - \varepsilon.$$

Note that the bias and variance defined here are those of the estimator, when estimating the (un)conditional error. These are quite different from the bias and variance of the learner itself. It is perfectly possible that a learner with high bias and low variance (say, linear regression) is evaluated using an estimator with low bias and high variance.

The variance of an estimator measures how much it varies around its own expected value; as such, it is independent of the estimand. Thus, the variance of any estimator considered here can be described independently of whether one wants to estimate, the conditional or the unconditional error. The bias and MSE, however, depend on which of these two one wants to estimate.

Most estimators considered in basic statistics, such as the sample mean, are deterministic functions: given a sample, the sample mean is uniquely determined. In that context,"variance" can only refer to the variance induced by the randomness of the sample; that is, a different sample would result in a different estimate, and the variance of these estimates is what the term "variance" refers to here.

The cross-validation estimator, however, is stochastic: It depends on random choices (typically some random partitioning or resampling $\pi$ of the data). Hence, even if the learner $L$ and sample $T$ are fixed, these estimators have a non-zero variance. In line with the literature, (**???**), we call this variance the *internal variance* of the estimator. It is the variance of the estimator over all possible partitionings or resamplings $\pi$ of the dataset $T$.

**Definition 13 (Internal variance of the (un)conditional error estimator)**

$$Var_\pi(\hat{\varepsilon}(L, T)) = \mathbf{E}_\pi[(\hat{\varepsilon} - \mathbf{E}_\pi[\hat{\varepsilon}])^2]. \tag{4}$$

The variance induced by the choice of the sample is then called the *sample variance*. Because the (un)conditional error estimator varies depending on the choice of the partitioning of the dataset, we first average over all possible partitionings of $T$ to obtain $\mathbf{E}_\pi[\hat{\varepsilon}_c]$:

**Definition 14 (Sample variance of the (un)conditional error estimator)**

$$Var_s(\hat{\varepsilon}_c) = Var_T(\mathbf{E}_\pi[\hat{\varepsilon}_c]) \tag{5}$$

We write the internal, sample, and total variance of an estimator $\hat{\varepsilon}$ as $\mathrm{Var}_\pi(\hat{\varepsilon})$ and $\mathrm{Var}_s(\hat{\varepsilon})$ and $\mathrm{Var}(\hat{\varepsilon})$, respectively. They are illustrated in figure **??**. As was also already noted by **?**, we can write the total variance of the estimator as follows by applying the law of total variance:

$$\mathrm{Var}(\hat{\varepsilon}) = \mathrm{Var}_s(\mathbf{E}_\pi[\hat{\varepsilon}]) + \mathbf{E}_T[\mathrm{Var}_\pi(\hat{\varepsilon})].$$

The concept of variance is not restricted to estimators only. We can also define the sample variance of the conditional errors $\varepsilon_c(L, T')$ over all datasets $T'$ of the same size as the given dataset $T$.

**Definition 15 (Sample variance of the conditional error)**

$$Var_s(\varepsilon_c) = Var_T(\varepsilon_c) \tag{6}$$

There is no reason to believe $\hat{\varepsilon}$ is an unbiased estimator for $\varepsilon_c(L, T)$. $\hat{\varepsilon}$ is based on a model learned from a dataset that is a subset of $T$, and therefore smaller; models learned from smaller datasets tend to be less accurate. The bias $B$ of $\hat{\varepsilon}$ is defined as:

**Definition 16 (Estimator bias)**

$$B(\hat{\varepsilon}) = \mathbf{E}_{T,\pi}[\hat{\varepsilon} - \varepsilon]. \tag{7}$$

The concepts defined in this section are illustrated in Figure **??**. It shows how different samples $T$ can be drawn from a population $P$. On each sample the conditional error $\varepsilon_c(L, T)$ can be computed. This gives rise to the sample variance of $\varepsilon_c$. On the same sample we can also compute a cross-validation estimate for $\varepsilon_c$ or $\varepsilon_u$. For this, multiple partitionings $\pi$ into folds are possible, where each partitioning results in a different estimate $\hat{\varepsilon}(L, T, \pi)$. Therefore, we say that the cross-validation estimator has internal variance. How the expected value of the cross-validation estimator over all possible partitionings of a sample $T$ varies, is expressed by the sample variance of the cross-validation estimator. This quantity is not necessarily equal to the sample variance of $\varepsilon_c$. Finally, we also note that $\mathbf{E}_\pi[\hat{\varepsilon}(L, T)]$ is not necessarily equal to $\varepsilon_c(L, T)$; the estimator may be biased.

## 4. Practical considerations for learner evaluation

### 4.1. Conditional error

The conditional error is defined on a single dataset, therefore, the estimator only has internal variance:

$$\mathrm{Var}(\hat{\varepsilon}_c) = \mathrm{Var}_\pi(\hat{\varepsilon}_c)$$

Internal variance is not a property of the learning problem, but of the resampling estimator itself. Therefore, decreasing it is beneficial because it improves the replicability of the experiment (**?**), and the quality of the estimator (**?**). In the case of cross-validation, this can be achieved by using *repeated* cross-validation. In fact, averaging over all possible partitionings of the dataset reduces the internal variance to zero.

However, this does not mean that the estimator converges to the true conditional error when the variance goes to zero; it has a bias, equal to:

$$B(\hat{\varepsilon}_c, \varepsilon_c) = \mathbf{E}_\pi[\hat{\varepsilon}_c] - \varepsilon_c$$

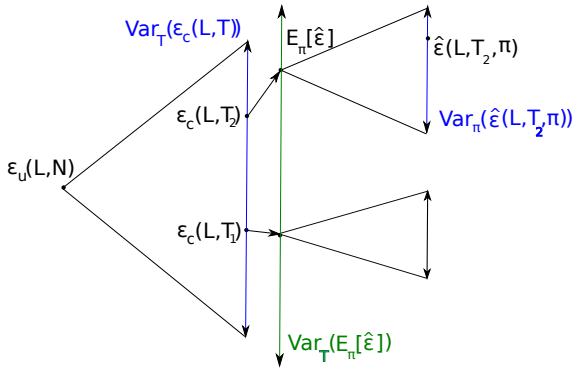



Figure 1: Illustration of the relationships between $\varepsilon_c$, $\varepsilon_u$, and the components of the mean squared error of the cross-validation estimator $\hat{\varepsilon}$ for these two population parameters.

.

A resampling estimator repeatedly splits the available dataset $T$ into a training and a test set. The model $m'$ that is learned on the training set generated by the estimator may differ from the model $m$ that is learned on the complete dataset $T$. Typically, the training set is smaller than $T$, and therefore systematically produces models with a larger error, making the estimator pessimistically biased with regard to the true conditional error.

When performing statistical inference, i.e., computing a confidence interval or applying a statistical test, a bias correction is therefore necessary. Unfortunately, the bias cannot readily be estimated, since we do not know the true conditional error.

### 4.2. Unconditional error

When interested in the expected performance of a learner on a random dataset sampled from the population, we are interested in the distributional properties of the conditional error, i.e., its expected value, $\varepsilon_u$, and its variance, $\mathrm{Var}(\varepsilon_c)$).

We already saw that the variance of the conditional error estimator equals:

$$\mathrm{Var}(\hat{\varepsilon}_c) = \mathrm{Var}_s(\mathbf{E}_\pi[\hat{\varepsilon}_c]) + \mathbf{E}_T[\mathrm{Var}_\pi(\hat{\varepsilon}_c)].$$

Often, it is not explicitly stated whether one is interested in estimating the conditional error, or the unconditional error. Moreover, regardless of which error one wants to estimate, only a single dataset is used to do it, although estimating $\mathrm{Var}(\hat{\varepsilon}_c)$ requires also estimating the sample variance of the conditional error. This fact is often ignored, and only $\mathrm{Var}_\pi(\hat{\varepsilon}_c)$ is estimated and used in a statistical test. Obviously, this results in incorrect inference results because the sample variance can be a significant component of $\mathrm{Var}(\hat{\varepsilon}_c)$ (**??**).

This problem is aggravated by repeated cross-validation. While it is recommended in a setting where one is interested in the performance of the actual *model*, it is not recommended when the goal is to do statistical inference about $\varepsilon_c$ or $\varepsilon_u$. The reduced internal variance,

combined with the absence of an estimate for the sample variance can lead to a significant underestimation of $\mathrm{Var}(\hat{\varepsilon}_c)$ and therefore a large probability of making a type I error, i.e., erroneously detecting a difference in performance between two learners.

But even if multiple samples are available, there is no consensus on how to properly estimate the variance of the cross-validation estimator. In fact, **?** proved that there does not exist an unbiased estimator for the variance of the cross-validation estimator, because the probability distribution of the estimator, $\mathcal{P}\hat{\varepsilon}_c(T, L, \pi)$, is not known exactly.

This means that the preferred statistical test to compare the performance of two learners by their cross-validation estimate is a debatable topic. Each statistical test has its own shortcomings. For instance, a well-known test is the binomial test for the difference in proportions, where the test statistic would be the average proportion of errors taken over all folds. This test assumes independence of the individual test errors on the instances, but this assumption is invalid, as the training sets generated by cross-validation partially overlap, and the errors computed on the same test fold result from the same model. Consequently, the test may have a larger probability of making a type I error than would be the case if the independence assumption was true.

An extension of this problem is the comparison of learners over multiple datasets. In this setting, we are again confronted with the problem of properly estimating the variance of the error estimates. However, the difficulty here is is not the lack of samples, or the dependencies between the error estimates; In his seminal work on this topic, **?** assumes that for each learner, an appropriate estimate of the unconditional error has been computed for each learner on each data population. Instead, the problem here is that the error estimates are computed on a number of datasets sampled from completely different populations, and therefore they are incommensurable.

## 5. Experiments

Our experiments try to answer the following questions:

- Does cross-validation estimate the conditional or the unconditional error?
- When comparing two models learned on a specific dataset, and ignoring statistical testing, how often does cross-validation correctly identify the model with the smallest prediction error?

### 5.1. Does the cross-validation estimator estimate the conditional or the unconditional error?

Our first experiment investigates whether the cross-validation estimator estimates the conditional error, the unconditional error, or neither. We do this by computing a cross-validation estimate $\hat{\varepsilon}(L, T)$ on a given dataset $T$, and comparing it to the true $\varepsilon_c$ and $\varepsilon_u$. These last two would normally not be available to the researcher, but we circumvent this problem by using a very large dataset $D$ as our population, so that we can compute all necessary quantities. The detailed experiment is described by Algorithm **??**:

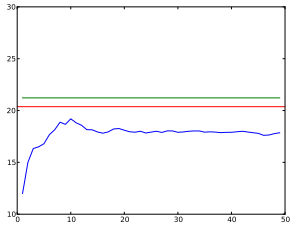
The experiment is repeated for different samples from $D$. $\varepsilon_u$ is computed as the mean of the conditional errors over all the samples. We follow the procedure that is often used in

**Algorithm 1:** Experimental procedure

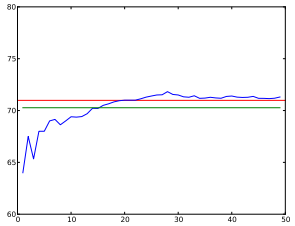
**Input**: A large dataset $D$ (population), learner $L$, and cross-validation estimator $CV$
$D$ is partitioned into a small dataset $T$ of $N$ instances and a large dataset $D \setminus T$.
We use $T$ to:
Compute $\varepsilon_c(L, T)$ by learning a model on $T$ and evaluating it on $D \setminus T$.
Compute a cross-validation estimate $\hat{\varepsilon}(L, T)$ and compare it to $\varepsilon_c$ and $\varepsilon_u$.

real experiments: We compute $\hat{\varepsilon}(L, T)$ on a single dataset from the population. The only variability of the estimator therefore arises from the random partitioning of the dataset. Using repeated cross-validation instead of regular cross-validation decreases this variance. When using an increasingly large number of repetitions, the estimator converges to an unknown value, which hopefully is $\varepsilon_c$ or $\varepsilon_u$, but this is to be investigated.

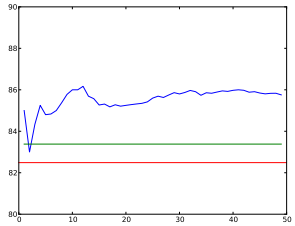
As our data populations, we use the following UCI datasets: Abalone, adult, king-rook versus king (kr-vs-k), mushroom, and nursery. The learning algorithms are: Naive Bayes (NB), nearest neighbors with 4 (4NN) and 10 neighbors (10NN), logistic regression (LR), the decision tree learner C4.5 (DT), and a Random Forest (RF). We also perform the experiment for a different number of folds of the cross-validation estimator, using 2-fold, 10-fold and 30-fold cross-validation. For every sample $T$ from $D$ we plot the model accuracy (blue) as computed by the repeated cross-validation estimator against the number of repetitions. The true conditional (green) and unconditional error (red) are also shown. Figure **??** presents a random selection of the results.



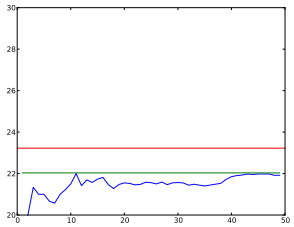
($a$) 4NN, kr-vs-k, 2fcv



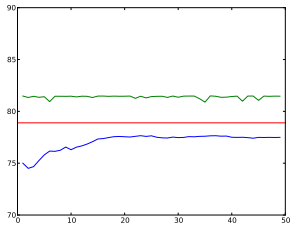
($b$) 10NN, nursery, 2fcv



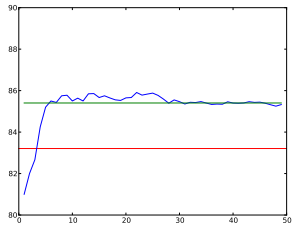
($c$) C4.5 on nursery, 10fcv



($d$) RF, kr-vs-k, 10fcv



($e$) LR, adult, 10fcv



($f$) NB, nursery, 2fcv

Figure 2: The horizontal axis shows the number of cross-validation repetitions. The vertical axis shows the the repeated cross-validation estimator for accuracy (blue), the conditional error [accuracy] (green), and the unconditional error [accuracy] (red).

The results demonstrate that repeated cross-validation indeed decreases the internal variance $\mathrm{Var}_\pi(\hat{\varepsilon})$ so that the estimate converges to $\mathbf{E}_\pi[\hat{\varepsilon}_c]$. However, we also see that this value is not equal to the conditional error, nor to the unconditional error. The estimator clearly has a bias, $\mathbf{E}_\pi[\hat{\varepsilon}_c] - \varepsilon$, which is different for every problem. Although, averaging over ten to twenty repetitions reduces this estimator bias. This is not true in every setting: The tenfold cross-validation estimate obtained for C4.5 on nursery, for instance, diverges from both $\varepsilon_c$ and $\varepsilon_u$. Another example is naive Bayes on nursery with twofold cross-validation. In this case, the estimator converges to the conditional error, but not the unconditional error.

### 5.2. Comparing learners with cross-validation

In the previous experiment we established that the cross-validation estimator computed on a single dataset is biased for both $\hat{\varepsilon}_c$ and $\hat{\varepsilon}_u$. However, if this bias is similar for every learner, two learning algorithms can still be compared by means of cross-validation. This is investigated in our next experiment. We focus only on estimating $\varepsilon_c$, but in the future we plan to extend our experiments to $\varepsilon_u$.

We again apply Algorithm **??**, but with one adjustment. Instead of one learner $L$, we apply step four and five on two learners $L_1$ and $L_2$, computing $\varepsilon_c$ and $\hat{\varepsilon}$ for both learners. We perform these computations for both learners on the same datasets $T$ with exactly the same settings for the cross-validation estimator. The resampling estimators are again 2-fold, 10-fold and 30-fold cross-validation, computed with 1 and 30 repetitions.

Based on the results for 100 samples from $D$, we construct a contingency table as follows, where we denote $\hat{\varepsilon}(L_i, T)$ as $\hat{\varepsilon}_i$ and $\varepsilon_c(L_i, T)$ as $\varepsilon_{c,i}$:

$M =$

| | $\varepsilon_{c,1} > \varepsilon_{c,2}$ | $\varepsilon_{c,1} \leq \varepsilon_{c,2}$ |
|---|---|---|
| $\hat{\varepsilon}_1 > \hat{\varepsilon}_2$ | | |
| $\hat{\varepsilon}_1 \leq \hat{\varepsilon}_2$ | | |

Our results are presented in Tables **??**, **??**, and **??** in the appendix [3]. As can be seen from these tables, 10-fold and 30-fold cross-validation outperform 2-fold cross-validation in detecting the winning model most often. Repeated cross-validation performs slightly better than regular cross-validation. This is consistent with the observations from our previous experiment, that repeated cross-validation often results in a more accurate estimate of $\varepsilon_c$ and $\varepsilon_u$.

It is interesting to see that when one learner is not clearly better than the other, cross-validation has difficulty selecting the winning model. Consider for instance the comparison of Naive Bayes and C4.5 on adult in Table **??**. $\varepsilon_c(C4.5)$ is smallest for more than half of the samples. However, for many samples where C4.5 wins, naive Bayes is selected by the cross-validation estimator as the winner. The opposite does not happen so often; on a sample where Naive Bayes wins, the cross-validation estimator often also selects Naive Bayes.

Let us focus on the case of 2-fold cross-validation for this problem, and compute the following conditional probabilities. By $L_1 > L_2$, we indicate that $L_1$ wins against $L_2$, i.e., the conditional error of $L_1$ is smaller than that of $L_2$.

3. Because of time restrictions, we were not able to obtain results for 30-fold cross-validation on kr-vs-k.

- $P(NB > DT | CV_{NB} > CV_{DT}) = 25/62 = 0.4$
- $P(NB \leq DT | CV_{NB} > CV_{DT}) = 37/62 = 0.6$
- $P(NB > DT | CV_{NB} \leq CV_{DT}) = 7/38 = 0.18$
- $P(NB \leq DT | CV_{NB} \leq CV_{DT}) = 31/38 = 0.82$

From these estimated probabilities we see that when the cross-validation estimator indicates that naive Bayes has a smaller conditional error, this is only true in 40% of cases. When cross-validation indicates that C4.5 wins, however, we have 82% certainty that this is indeed true. The reason is perhaps that overall, C4.5 wins on most samples: 68 out of 100 samples. Therefore, changes made by the cross-validation estimator in the sample will most likely create a sample for which the decision tree wins.

We can also compute the opposite conditional probabilities:

- $P(CV_{NB} > CV_{DT} | NB > DT) = 25/32 = 0.78$
- $P(CV_{NB} \leq CV_{DT} | NB > DT) = 7/32 = 0.22$
- $P(CV_{NB} > CV_{DT} | NB \leq DT) = 37/68 = 0.54$
- $P(CV_{NB} \leq CV_{DT} | NB \leq DT) = 31/68 = 0.56$

We see that when we select a sample on which we know naive Bayes wins, it is 78% certain that cross-validation will detect this. However, when we select a sample for which we know the decision tree wins, the cross-validation estimate is no better than a random guess (50% probability).

Another example is the comparison of naive Bayes and the random forest with 2-fold cross-validation on kr-vs-k (Table **??**). Here, the random forest wins on more than half of the samples. The estimated conditional probabilities are as follows:

- $P(NB > RF | CV_{NB} > CV_{RF}) = 2/5 = 0.4$
- $P(NB \leq RF | CV_{NB} > CV_{RF}) = 3/5 = 0.6$
- $P(NB > RF | CV_{NB} \leq CV_{RF}) = 33/95 = 0.35$
- $P(NB \leq RF | CV_{NB} \leq CV_{RF}) = 62/95 = 0.65$

Again, on a sample where the random forest wins, the probability that the same conclusion is reached by the cross-validation estimator is larger than 0.5, while the opposite is true when naive Bayes wins.

- $P(CV_{NB} > CV_{RF} | NB > RF) = 2/40 = 0.05$
- $P(CV_{NB} \leq CV_{RF} | NB > RF) = 38/40 = 0.95$
- $P(CV_{NB} > CV_{RF} | NB \leq RF) = 3/65 = 0.05$
- $P(CV_{NB} \leq CV_{RF} | NB \leq RF) = 62/65 = 0.95$

Here, the results are even more extreme than in the previous example. Regardless of whether a sample is selected for which we know the random forest wins, or naive Bayes, the cross-validation estimator concludes with high probability that the random forest wins.

## 6. Conclusions

This paper discusses a number of crucial points to take into account when estimating the error of a predictive model with cross-validation. It is motivated by the observation that, although being an essential task in machine learning research, there does not seem to be a consensus on how to perform this task.

Our first point is that a researcher should always be clear on whether they are estimating the error of a *model*, i.e., the conditional error, or that of *learner*, i.e., the unconditional error. Estimating one or the other requires a different approach. In machine learning research, most often the relevant quantity is the error of the learner. This involves estimating the expected value of the conditional error, and its variance over different samples from the population. By definition, these quantities cannot be estimated on a single sample. This is in contrast with what is often observed in practice: Although the context of the paper suggests that the researcher in interested in the learner, the experiments are set up as if they were interested in the model learned on the single available dataset.

Our experiments show that when using cross-validation for choosing between two models, the best performing model is not always chosen. The standard approach for handling the uncertainty of the outcome introduced by selecting a single sample, and partitioning that into folds, is to use statistical testing. However, in this particular situation, there are two problems with this approach.

First, statistical testing requires an accurate estimate of the variance of the cross-validation estimate of the prediction error. Unfortunately, a wealth of statistical tests exist and often it is not clear for a researcher which one to use. In fact, estimating the variance of the cross-validation estimator is notoriously difficult because of dependencies between the individual test errors, and no unanimously recommended statistical test exists for the task.

Second, often only a single dataset is available from the population for which one wants to know learner performance. This means that the obtained variance estimate does not account for sample variance of the cross-validation estimate. Instead, the variance estimate only accounts for the variance of the error estimates on different folds, i.e., the internal variance. This internal variance is a component of the total variance of the cross-validation estimate, rather than a substitute for the sample variance.

The internal variance will even be zero when performing estimation with repeated cross-validation with a sufficient number of repetitions. This is because the internal variance is a property of the estimation method (cross-validation), and not of the test statistic. Therefore, having low internal variance means having a more reliable error estimate. Our experiments indicate that in most cases, ten to twenty repetitions indeed lead to a more accurate error estimate than performing no repetitions, both for the conditional and the unconditional error.

However, repeated cross-validation is of no advantage when performing statistical inference, as the internal variance is no substitute for the sample variance of the estimator. Moreover, if the sample variance is indeed substituted by internal variance, a performance difference between two learners can always be detected by using a sufficiently large number of repetitions.

This discussion leads us to question the usefulness of statistical testing in the context of evaluating predictive models with cross-validation. We advocate instead to always provide a clear interpretation of the experimental results. For instance, by clearly stating whether one is estimating the conditional or the unconditional error.

## Appendix A. Contingency tables

| $L_1, L_2$ | folds | 2 | 10 | 30 |
|---|---|---|---|---|
| NB-DT | adult | $\left[\begin{smallmatrix}25 & 37\\ 7 & 31\end{smallmatrix}\right], \left[\begin{smallmatrix}14 & 48\\ 1 & 37\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}32 & 30\\ 7 & 31\end{smallmatrix}\right], \left[\begin{smallmatrix}28 & 34\\ 5 & 33\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}32 & 30\\ 6 & 32\end{smallmatrix}\right], \left[\begin{smallmatrix}37 & 25\\ 4 & 34\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}8 & 8\\ 36 & 48\end{smallmatrix}\right], \left[\begin{smallmatrix}8 & 8\\ 30 & 54\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}8 & 8\\ 29 & 55\end{smallmatrix}\right], \left[\begin{smallmatrix}8 & 8\\ 15 & 69\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}53 & 12\\ 24 & 11\end{smallmatrix}\right], \left[\begin{smallmatrix}61 & 4\\ 27 & 8\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}52 & 13\\ 13 & 22\end{smallmatrix}\right], \left[\begin{smallmatrix}52 & 13\\ 11 & 24\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}50 & 15\\ 12 & 23\end{smallmatrix}\right], \left[\begin{smallmatrix}49 & 16\\ 9 & 26\end{smallmatrix}\right]$ |
| NB-4NN | adult | $\left[\begin{smallmatrix}51 & 24\\ 5 & 20\end{smallmatrix}\right], \left[\begin{smallmatrix}50 & 25\\ 3 & 22\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}67 & 8\\ 1 & 24\end{smallmatrix}\right], \left[\begin{smallmatrix}69 & 6\\ 1 & 24\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}70 & 5\\ 1 & 24\end{smallmatrix}\right], \left[\begin{smallmatrix}71 & 4\\ 1 & 24\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}22 & 15\\ 30 & 33\end{smallmatrix}\right], \left[\begin{smallmatrix}20 & 17\\ 34 & 29\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}19 & 18\\ 20 & 43\end{smallmatrix}\right], \left[\begin{smallmatrix}17 & 20\\ 19 & 44\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}99 & 1\\ 0 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}100 & 0\\ 0 & 0\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}97 & 3\\ 0 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}100 & 0\\ 0 & 0\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}96 & 4\\ 0 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}98 & 2\\ 0 & 0\end{smallmatrix}\right]$ |
| NB-10NN | adult | $\left[\begin{smallmatrix}34 & 34\\ 6 & 26\end{smallmatrix}\right], \left[\begin{smallmatrix}26 & 42\\ 6 & 26\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}47 & 21\\ 3 & 29\end{smallmatrix}\right], \left[\begin{smallmatrix}51 & 17\\ 5 & 27\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}59 & 9\\ 4 & 28\end{smallmatrix}\right], \left[\begin{smallmatrix}57 & 11\\ 1 & 31\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}23 & 16\\ 38 & 23\end{smallmatrix}\right], \left[\begin{smallmatrix}20 & 19\\ 36 & 25\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}12 & 27\\ 21 & 40\end{smallmatrix}\right], \left[\begin{smallmatrix}15 & 24\\ 18 & 43\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}99 & 0\\ 1 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}99 & 0\\ 1 & 0\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}95 & 4\\ 1 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}97 & 2\\ 1 & 0\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}94 & 5\\ 1 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}95 & 4\\ 1 & 0\end{smallmatrix}\right]$ |
| NB-LR | adult | $\left[\begin{smallmatrix}9 & 25\\ 15 & 51\end{smallmatrix}\right], \left[\begin{smallmatrix}4 & 30\\ 4 & 62\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}13 & 21\\ 12 & 54\end{smallmatrix}\right], \left[\begin{smallmatrix}12 & 22\\ 12 & 54\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}15 & 19\\ 13 & 53\end{smallmatrix}\right], \left[\begin{smallmatrix}19 & 15\\ 15 & 51\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}10 & 28\\ 21 & 41\end{smallmatrix}\right], \left[\begin{smallmatrix}10 & 28\\ 16 & 46\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}18 & 20\\ 14 & 48\end{smallmatrix}\right], \left[\begin{smallmatrix}16 & 22\\ 11 & 51\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}27 & 4\\ 55 & 14\end{smallmatrix}\right], \left[\begin{smallmatrix}30 & 1\\ 65 & 4\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}15 & 16\\ 23 & 46\end{smallmatrix}\right], \left[\begin{smallmatrix}21 & 10\\ 19 & 50\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}14 & 17\\ 16 & 53\end{smallmatrix}\right], \left[\begin{smallmatrix}18 & 13\\ 16 & 53\end{smallmatrix}\right]$ |
| NB-RF | adult | $\left[\begin{smallmatrix}5 & 30\\ 8 & 57\end{smallmatrix}\right], \left[\begin{smallmatrix}2 & 33\\ 4 & 61\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}12 & 23\\ 13 & 52\end{smallmatrix}\right], \left[\begin{smallmatrix}8 & 27\\ 9 & 56\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}11 & 24\\ 15 & 50\end{smallmatrix}\right], \left[\begin{smallmatrix}14 & 21\\ 11 & 54\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}2 & 3\\ 33 & 62\end{smallmatrix}\right], \left[\begin{smallmatrix}2 & 3\\ 19 & 76\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}3 & 2\\ 20 & 75\end{smallmatrix}\right], \left[\begin{smallmatrix}3 & 2\\ 18 & 77\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}42 & 16\\ 25 & 17\end{smallmatrix}\right], \left[\begin{smallmatrix}52 & 6\\ 32 & 10\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}47 & 11\\ 14 & 28\end{smallmatrix}\right], \left[\begin{smallmatrix}48 & 10\\ 12 & 30\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}42 & 16\\ 10 & 32\end{smallmatrix}\right], \left[\begin{smallmatrix}48 & 10\\ 10 & 32\end{smallmatrix}\right]$ |

Table 1: Contingency tables for the comparison of NB with DT, 4NN, 10NN, LR, and RF.

| $L_1, L_2$ | folds | 2 | 10 | 30 |
|---|---|---|---|---|
| DT-4NN | adult | $\left[\begin{smallmatrix}84 & 16\\ 0 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}98 & 2\\ 0 & 0\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}91 & 9\\ 0 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}96 & 4\\ 0 & 0\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}91 & 9\\ 0 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}91 & 9\\ 0 & 0\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}50 & 41\\ 4 & 5\end{smallmatrix}\right], \left[\begin{smallmatrix}56 & 35\\ 6 & 3\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}55 & 36\\ 3 & 6\end{smallmatrix}\right], \left[\begin{smallmatrix}55 & 36\\ 5 & 4\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}95 & 5\\ 0 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}100 & 0\\ 0 & 0\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}95 & 5\\ 0 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}100 & 0\\ 0 & 0\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}96 & 4\\ 0 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}99 & 1\\ 0 & 0\end{smallmatrix}\right]$ |
| DT-10NN | adult | $\left[\begin{smallmatrix}50 & 38\\ 4 & 8\end{smallmatrix}\right], \left[\begin{smallmatrix}60 & 28\\ 3 & 9\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}64 & 24\\ 6 & 6\end{smallmatrix}\right], \left[\begin{smallmatrix}64 & 24\\ 6 & 6\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}62 & 26\\ 5 & 7\end{smallmatrix}\right], \left[\begin{smallmatrix}61 & 27\\ 6 & 6\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}57 & 33\\ 6 & 4\end{smallmatrix}\right], \left[\begin{smallmatrix}54 & 36\\ 5 & 5\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}45 & 45\\ 6 & 4\end{smallmatrix}\right], \left[\begin{smallmatrix}44 & 46\\ 6 & 4\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}96 & 4\\ 0 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}100 & 0\\ 0 & 0\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}93 & 7\\ 0 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}99 & 1\\ 0 & 0\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}94 & 6\\ 0 & 0\end{smallmatrix}\right], \left[\begin{smallmatrix}98 & 2\\ 0 & 0\end{smallmatrix}\right]$ |
| DT-LR | adult | $\left[\begin{smallmatrix}1 & 7\\ 28 & 64\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 8\\ 21 & 71\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}1 & 7\\ 31 & 61\end{smallmatrix}\right], \left[\begin{smallmatrix}2 & 6\\ 26 & 66\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}4 & 4\\ 26 & 66\end{smallmatrix}\right], \left[\begin{smallmatrix}3 & 5\\ 26 & 66\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}33 & 50\\ 6 & 11\end{smallmatrix}\right], \left[\begin{smallmatrix}23 & 60\\ 7 & 10\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}37 & 46\\ 9 & 8\end{smallmatrix}\right], \left[\begin{smallmatrix}36 & 47\\ 8 & 9\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}4 & 5\\ 39 & 52\end{smallmatrix}\right], \left[\begin{smallmatrix}6 & 3\\ 40 & 51\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}4 & 5\\ 15 & 76\end{smallmatrix}\right], \left[\begin{smallmatrix}4 & 5\\ 16 & 75\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}3 & 6\\ 17 & 74\end{smallmatrix}\right], \left[\begin{smallmatrix}5 & 4\\ 22 & 69\end{smallmatrix}\right]$ |
| DT-RF | adult | $\left[\begin{smallmatrix}2 & 5\\ 14 & 79\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 7\\ 2 & 91\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}2 & 5\\ 29 & 64\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 7\\ 15 & 78\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}2 & 5\\ 24 & 69\end{smallmatrix}\right], \left[\begin{smallmatrix}1 & 6\\ 20 & 73\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}11 & 18\\ 28 & 43\end{smallmatrix}\right], \left[\begin{smallmatrix}6 & 23\\ 8 & 63\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}13 & 16\\ 29 & 42\end{smallmatrix}\right], \left[\begin{smallmatrix}9 & 20\\ 21 & 50\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}7 & 26\\ 24 & 43\end{smallmatrix}\right], \left[\begin{smallmatrix}5 & 28\\ 7 & 60\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}13 & 20\\ 24 & 43\end{smallmatrix}\right], \left[\begin{smallmatrix}14 & 19\\ 15 & 52\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}15 & 18\\ 23 & 44\end{smallmatrix}\right], \left[\begin{smallmatrix}18 & 15\\ 22 & 45\end{smallmatrix}\right]$ |

Table 2: Contingency tables for the comparison DT with 4NN, 10NN, LR, and RF.

| $L_1, L_2$ | folds | 2 | 10 | 30 |
|---|---|---|---|---|
| 4NN-10NN | adult | $\left[\begin{smallmatrix}0 & 1\\ 7 & 92\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 1\\ 0 & 99\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 1\\ 11 & 88\end{smallmatrix}\right], \left[\begin{smallmatrix}1 & 0\\ 7 & 92\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}1 & 0\\ 7 & 92\end{smallmatrix}\right], \left[\begin{smallmatrix}1 & 0\\ 8 & 91\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}32 & 30\\ 21 & 17\end{smallmatrix}\right], \left[\begin{smallmatrix}33 & 29\\ 21 & 17\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}27 & 35\\ 18 & 20\end{smallmatrix}\right], \left[\begin{smallmatrix}24 & 38\\ 12 & 26\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}10 & 15\\ 31 & 44\end{smallmatrix}\right], \left[\begin{smallmatrix}15 & 10\\ 30 & 45\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}11 & 14\\ 23 & 52\end{smallmatrix}\right], \left[\begin{smallmatrix}11 & 14\\ 23 & 52\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}13 & 12\\ 22 & 53\end{smallmatrix}\right], \left[\begin{smallmatrix}12 & 13\\ 24 & 51\end{smallmatrix}\right]$ |
| 4NN-LR | adult | $\left[\begin{smallmatrix}0 & 0\\ 8 & 92\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 0\\ 2 & 98\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 1 & 99\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 0\\ 3 & 97\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 3 & 97\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}18 & 25\\ 15 & 42\end{smallmatrix}\right], \left[\begin{smallmatrix}19 & 24\\ 9 & 48\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}21 & 22\\ 21 & 36\end{smallmatrix}\right], \left[\begin{smallmatrix}22 & 21\\ 16 & 41\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}0 & 0\\ 2 & 98\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right]$ |
| 4NN-RF | adult | $\left[\begin{smallmatrix}0 & 0\\ 2 & 98\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 0\\ 3 & 97\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 1 & 99\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 0\\ 3 & 97\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 2 & 98\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}0 & 1\\ 30 & 69\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 1\\ 25 & 74\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 1\\ 32 & 67\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 1\\ 26 & 73\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}0 & 0\\ 1 & 99\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right]$ |
| 10NN-LR | adult | $\left[\begin{smallmatrix}1 & 1\\ 25 & 73\end{smallmatrix}\right], \left[\begin{smallmatrix}2 & 0\\ 18 & 80\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}2 & 0\\ 13 & 85\end{smallmatrix}\right], \left[\begin{smallmatrix}2 & 0\\ 8 & 90\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}2 & 0\\ 16 & 82\end{smallmatrix}\right], \left[\begin{smallmatrix}2 & 0\\ 13 & 85\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}8 & 30\\ 20 & 42\end{smallmatrix}\right], \left[\begin{smallmatrix}13 & 25\\ 10 & 52\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}21 & 17\\ 25 & 37\end{smallmatrix}\right], \left[\begin{smallmatrix}24 & 14\\ 21 & 41\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right]$ |
| 10NN-RF | adult | $\left[\begin{smallmatrix}0 & 1\\ 19 & 80\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 1\\ 14 & 85\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}1 & 0\\ 11 & 88\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 1\\ 10 & 89\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 1\\ 18 & 81\end{smallmatrix}\right], \left[\begin{smallmatrix}1 & 0\\ 16 & 83\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}0 & 0\\ 25 & 75\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 18 & 82\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 0\\ 43 & 57\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 36 & 64\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}0 & 0\\ 1 & 99\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 0\\ 2 & 98\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}0 & 0\\ 2 & 98\end{smallmatrix}\right], \left[\begin{smallmatrix}0 & 0\\ 0 & 100\end{smallmatrix}\right]$ |
| LR-RF | adult | $\left[\begin{smallmatrix}18 & 37\\ 15 & 30\end{smallmatrix}\right], \left[\begin{smallmatrix}14 & 41\\ 19 & 26\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}25 & 30\\ 24 & 21\end{smallmatrix}\right], \left[\begin{smallmatrix}24 & 31\\ 26 & 19\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}28 & 27\\ 25 & 20\end{smallmatrix}\right], \left[\begin{smallmatrix}28 & 27\\ 28 & 17\end{smallmatrix}\right]$ |
| | kr-vs-k | $\left[\begin{smallmatrix}2 & 2\\ 47 & 49\end{smallmatrix}\right], \left[\begin{smallmatrix}2 & 2\\ 43 & 53\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}1 & 3\\ 40 & 56\end{smallmatrix}\right], \left[\begin{smallmatrix}2 & 2\\ 32 & 64\end{smallmatrix}\right]$ | |
| | nursery | $\left[\begin{smallmatrix}28 & 55\\ 5 & 12\end{smallmatrix}\right], \left[\begin{smallmatrix}24 & 59\\ 3 & 14\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}61 & 22\\ 9 & 8\end{smallmatrix}\right], \left[\begin{smallmatrix}66 & 17\\ 12 & 5\end{smallmatrix}\right]$ | $\left[\begin{smallmatrix}57 & 26\\ 12 & 5\end{smallmatrix}\right], \left[\begin{smallmatrix}70 & 13\\ 12 & 5\end{smallmatrix}\right]$ |



Table 3: Contingency tables for the comparison of 4NN, 10NN, LR, and RF.